针对互联网数据的新闻转载引用分析

摘　要：互联网、大数据和新媒体技术的发展带来媒体传播渠道和内容形态革命性变化，分析新闻在不同渠道媒体采用和传播情况是构建大数据驱动采编和传播决策的重要组成部分，对于提升通讯社国内和国际传播能力具有十分重要的意义。然而，由于互联网和新媒体数据格式不规范、转载和引用不注明来源等问题，新媒体的新闻转载引用分析难度大。本文从多源头收集网站、电子报纸、微信公众号、移动客户端等新闻数据，覆盖全球5000余家中英文媒体、40余万个新媒体账户。利用信息智能比对技术，跟踪新闻在全媒体的落地采用，构建新闻转载和引用分析系统，为进一步分析媒体传播路径，掌握国内外媒体传播规律，提升国内外舆论传播力奠定了基础。文中介绍了新闻转载引用分析的工作原理和建设意义，对关键技术实现进行了深入研究，在此基础上提出了新闻转载引用分析未来的发展建议。

关键词：新闻转载引用；文本相似度大数据；Hadoop Spark

中图分类号：TP392　　**文献标识码：**A

文章编号：1671-0134（2017）11-089-03　　**DOI：**10.19483/j.cnki.11-4653/n.2017.11.029

文／陈辛夷　陈 珺　王 熠

引言

互联网、大数据和新媒体技术的发展带来媒体传播渠道和内容形态革命性变化。如何利用智能分析技术，在互联网大数据中定位和跟踪新闻转载和引用的信息，及时反映新闻被国内外媒体采用的情况，是构建大数据驱动采编和传播决策的重要组成部分，对于提升通讯社国内和国际传播能力具有十分重要的意义。

本文从多源头收集网站、电子报纸、微信公众号、移动客户端等数据，覆盖全球5000余家中英文媒体、40余万个新媒体账户，利用信息智能比对技术，跟踪新闻在全媒体的落地采用，构建新闻转载和引用分析系统，为进一步分析媒体传播路径，掌握国内外媒体传播规律，提升国内外舆论传播力奠定了基础。

1. 新闻转载引用分析概念

新闻转载引用分析是针对一篇原创新闻，通过一系列技术手段分析海量实时的互联网大数据，识别出其中转载和引用该新闻的媒体。

转载指报刊或网站等媒体上发布其他媒体已经发表过的新闻。在对内报道中，新闻被媒体全文转载的情况比较常见。

引用指报刊或网站等媒体的文章中部分引用了其他媒体已经发表过的新闻中的语句或信息。在对外报道中，海外媒体特别是国际主流媒体通常引用新闻中的一段或一句，或者将原文中的信息转述表达。在新闻报道中，引用的场景一种是引述事实再展开深入报道；另一种是引述观点进而阐述相同或相反的观点。

显性转载引用指报刊或网站在转载或引用时注明转载或引用媒体的情况。一种情况是在转载时保留电头；另一种情况是在引用时指明“据某媒体报道”。

隐性转载引用在新闻的转载引用中存在文章中不注明来源的情况，称为隐性转载或引用。与显性转载引用相比，隐性转载引用的识别难度更大。随着互联网技术的发展，各种新媒体不断涌现，在拓展传播边界的同时也存在着转载不规范的问题。

2. 新闻转载引用分析的意义

通过分析新闻在中英文网站、电子报纸、微信和移动客户端的转载和引用情况，标记引用的段落和句子，识别采用媒体、采用时间和采用的版面等信息，可以及时追踪和分析新闻被全媒体采用的情况，进而可以统计和评估采编人员的工作成果，并对稿件的传播效果进行分析，为指导进一步提高新闻传播影响力提供数据支持。

3. 新闻转载引用分析工作原理

本文提出一种基于文本语义对比进行新闻转载引用分析的技术，主要包含新闻特征提取、相似新闻聚类、新闻转载引用关系判定、结果校验几个步骤。

新闻特征提取：采用网页信息抽取技术提取互联网新闻数据特征。对每篇稿件，通过分析网页的结构，使用机器学习与规则相融合的算法抽取出该新闻的发布时间。

相似新闻聚类：使用相似簇划分算法对采集的互联网新闻数据按照语义相似度进行划分，每个相似簇内部的新闻都是语义相似的，这些新闻数据之间可能存在隐式转载的关系。

新闻转载引用关系判定：综合相似簇内新闻间的相似度和新闻的发布时间等信息，根据经验判定阈值，对新闻的转载引用关系进行分析判定，得出新闻间的转载引用关系。

结果校验：对判定结果进行再次校验。

4. 新闻转载引用分析技术原理

系统总体数据处理架构如图1所示。主要架构设计思路和数据处理过程分为以下几个部分：

数据引进层：通过大规模数据采集和第三方引入的互联网新闻数据，首先使用Redis进行排重，然后进行数据的预处理及ETL，形成规则数据，得到结构化数据。

任务调度层：基于 Kafka 分布式消息队列，实现互联网数据的接入和缓冲。对 Kafka 消息队列里的数据结合实时 Spark Steaming 流式计算和离线大规模 M/R 计算框架进行新闻转载引用分析。

数据存储层：面对海量新闻数据，分布式存储可以实现高效的业务逻辑运算、可伸缩的存储部署策略和高可用的冗余式存储。MySQL 作为转载引用统计结果的基础存储数据库，负责数据模型的定义与数据积累，但不对外提供复杂的查询服务。ElasticSearch 首先作为 MySQL 核心业务表的镜像进行数据同步，同时实现多表关联和数据冗余，提升查询性能。其次，作为数据服务业务的实时服务端，提供数据服务的在线查询。Hive 作为数据服务的离线服务端，提供离线的大规模数据查询分析服务。FastDFS 作为离散文件的存储系统，提供图片、PDF 和报告 Excel 文件的存储管理。

集成服务层：针对业务需求，依托服务总线技术将底层数据通过灵活多样的查询和数据提取逻辑发布至上层服务接口，实现对外的通用服务接口。基于 Zookeeper 和 Dubbo 实现服务总线，统一协调调度，统一配置管理。

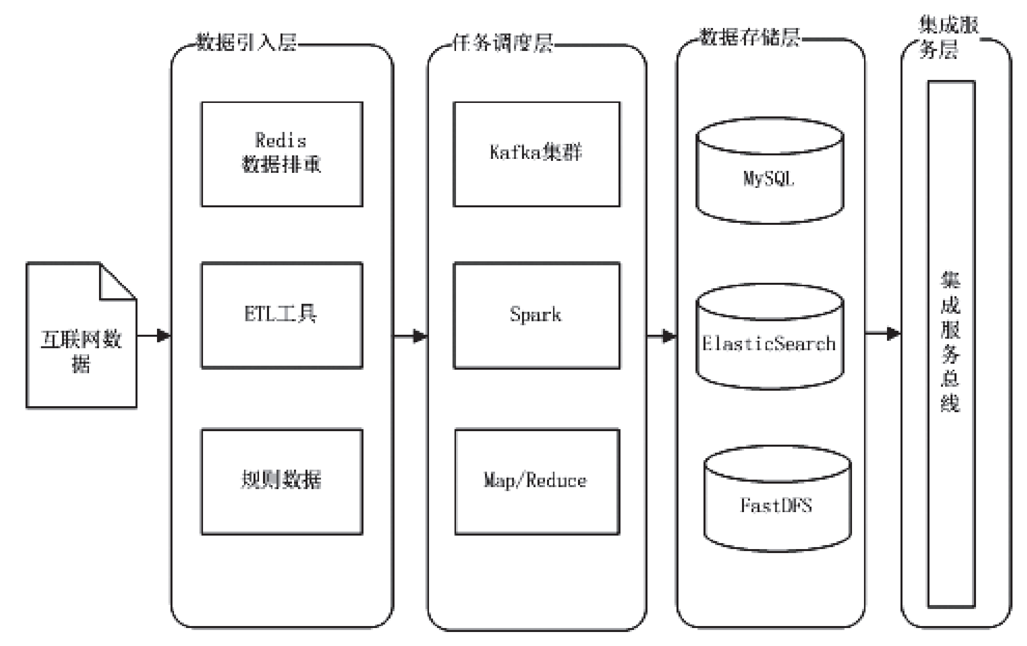


图 1　系统数据处理架构设计

5. 新闻转载引用分析关键技术

5.1 网页信息抽取

从网页源码中解析内容信息，传统的方法一般会采用递归解析子标签的方式，逐一获取标签内容。但在实际应用中，该方式在解析复杂的网页源码时，复杂度过高，消耗的资源过大。为解决这种问题，本文设计网页内容解析算法，采用 XPATH 技术与网页结构树递归解析结合的方式抽取网页内容。XPath 即为 XML 路径语言，它是一种用来确定 XML 文档中某部分位置的语言，它提供在数据结构树中找寻节点的能力。

网页的主体内容信息一般都在特定的 HTML 标签或者其子标签下，本算法先通过 XPATH 技术获取网页中的主体正文块，对于每一个正文块，构造网页结构树，在结构树上以递归的方式遍历全部的标签。在递归处理过程中，以标签全路径来记录遍历过的路径，避免标签被重复解析。在算法遍历的过程中，可以获取网页所包含的标题、正文、网页链接、来源、发布时间等信息。

5.2 文本相似度比对

使用文本相似度比对算法，将文本划分为不同的相似簇。本文使用经典的 VSM（向量空间模型）与 Bag of Words(BOW) 作为文档表示模型，该模型的基本思想是将文档分为若干的特征项，通过对特征项权重的量化计算进而将整个文档用一特征项的权重为分量的向量来表示，在将文档用特征向量的方式表示为数学模型后，再基于特征向量进行文档间的相似度计算。使用 TF-IDF 算法作为特征项的权重值。文本相似度计算的流程如图 2 所示。

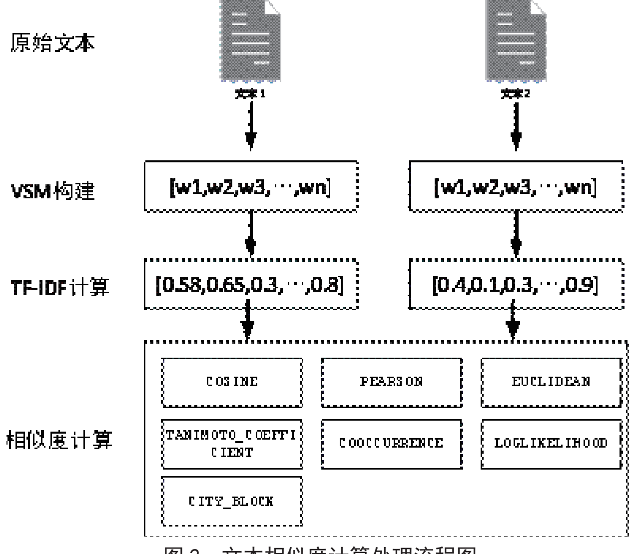


图 2　文本相似度计算处理流程图

在文本相似性比对算法中，文本相似度量算法扮演了

重要的角色，常用的相似度度量方法有：皮尔逊相关系数（Pearson Correlation Coefficient ,PCC）、余弦相似度（Cosine Similarity）、欧几里得相似度（Euclidean Similarity）等，经对比发现，皮尔逊相关系数更适合本算法。皮尔逊相关系数是计算两个向量线性相关度的一个指标，其计算公式如下：

$$\rho_{X,Y}=\frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2\sum(Y-\overline{Y})^2}}$$

5.3 新闻转载引用关系构建

基于文本相似簇，再利用网页新闻的来源、发布时间等信息，实现转载引用网络的构建。本文使用了图数据库构建与存储转载引用网络，支持数据的动态更新和多级转载引用关系的查询。最终，利用网络路径追踪技术，可以追溯每一篇新闻的转载引用路径，定位追踪新闻的转载引用情况。

6. 相似文本簇划分的具体技术实现

针对不同的应用场景开发了两套相似文本簇划分系统，分别是适合批处理的基于 Hadoop 平台的相似文本簇划分和适合实时计算的基于分布式内存实时计算的相似文本簇划分。

6.1 基于 Hadoop 平台的相似文本簇划分

Hadoop 作为大数据处理领域最成熟的解决方案，其以分布式文件系统 HDFS 和分布式计算模型 MapReduce 为代表的技术在大数据批处理领域取得了很大的成功。此外 Hadoop 拥有完善的生态系统，可以提供丰富的组件支持，本文使用了数据挖掘工具包 Mahout 中的一些算法，极大地简化了处理的难度。

6.2 基于分布式内存实时计算的相似文本簇划分

基于分布式内存实时计算的相似文本簇划分系统主要针对一些对实时性要求比较高的场景。该系统可以实现亚秒级响应的数据处理，处理框架图如图 3。

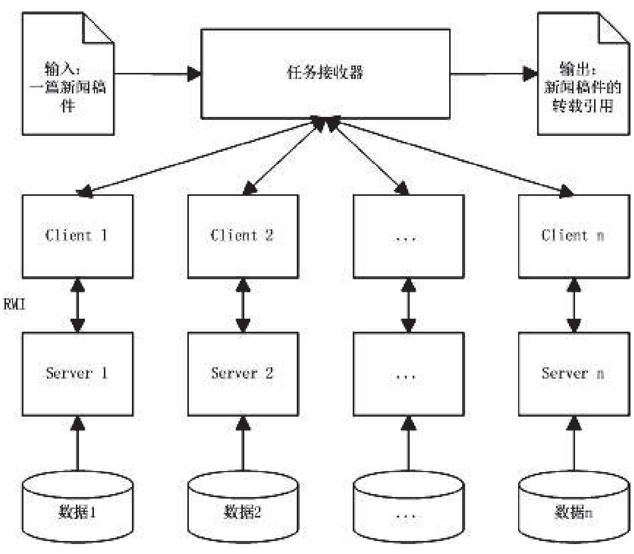


图 3　基于分布式内存实时计算的相似文本簇划分处理框架图

7. 集成及测试效果

经过多轮测试和算法优化，目前中文文字新闻转载引用分析准确率达到 95% 以上，英文文字新闻转载引用分析准确率达到 90% 以上。

8. 面向互联网大数据的新闻转载引用分析的应用展望

传播路径分析结合相似文本簇划分对新闻的整个传播路径进行分析，找到传播路径中的关键媒体或新媒体账户。

专题报道分析针对专题报道中的一组新闻进行转载和引用分析，结合专题的时间、地域、事件发展过程等分析总结其中的传播规律。

舆论引导力分析在一个新闻事件的报道中，通过分析某一篇新闻前后的新闻报道，研究这篇新闻起到了怎样的舆论引导作用，达成了怎样的效果。

结语

2017 年 4 月，系统上线试运行，提供全社采编人员实时查询稿件在全媒体的采用情况，提供总社和分社新闻采编业务统计数据和新闻采编人员考核数据的基础数据，提供全社全媒体报道发稿、采用和互动情况的大屏展示，初步取得了较好的效果。随着应用的不断深入，采编人员和统计人员都对系统提出了新的要求。系统会继续针对图片视频等多媒体稿件的采用分析、小语种稿件的采用分析等难点课题进行进一步研究。媒

参考文献

[1] Holden Karau 等 . Spark 快速大数据分析 [J]. 北京：人民邮电出版社，2015（10）：161-185.

[2] Sean Owen 等 .Mahout 实战 [J]. 北京：人民邮电出版社，2014（3）：40-47.

[3] Tom White.Hadoop 权威指南 [J]. 北京：清华大学出版社，2011 年（7）：160-174.

（作者单位：新华通讯社通信技术局）